R404: Applied Econometrics II

Topic 1: The Classical Linear Regression Model: OLS Estimation (in Matrix Form)

Kaloyan Ganev

2022/2023

Lecture Contents

Introduction

Introduction

- The goal of this course is to expand your econometrics knowledge and skills
- We will rely on the concepts and methods learned so far in the first programme module
- Yet, some complications are necessary and they will be introduced accordingly
- We start with an review of convergence concepts as they are frequently needed

Introduction (2)

- Then we also review OLS estimation so that we can make a smooth transition between courses
- In addition to the above, we aim to use this lecture as a tool facilitating your adaptation to matrix notation
- The text that will follow most closely is Greene's (2013), although the 2018 edition can also be used
- When necessary, other readings will also supplement it

A review of stochastic convergence

Convergence in Distribution

- The weakest form of stochastic convergence, also known as **weak convergence**
- Formal definition:

Definition 1 (Convergence in distribution)

Let $\{X_i\}$ be a sequence of real-valued random variables. This sequence converges in distribution to a random variable X if:

$$\lim_{n \to \infty} F_n(x) = F(x),$$

$\forall x \in \mathbb{R}$ at which F is continuous.

- *Interpretation:* new outcomes are increasingly better modelled by a given probability distribution
- Notation:

$$X_n \xrightarrow{d} X$$

Convergence in Probability

Definition 2 (Convergence in probability)

A random sequence $\{X_i\}$ converges in probability to the random variable X if $\forall \varepsilon > 0$:

$$\lim_{n\to\infty} Prob(|X_n - X| \geq \varepsilon) = 0$$

- *Interpretation:* the probability of unusual outcomes becomes very small when n gets large
- Notation:

$$\text{plim}_{n\to\infty} X_n = X \quad \text{or} \quad X_n \xrightarrow{p} X$$

- Convergence in probability implies convergence in distribution; the implication in the reverse direction is only valid when the limiting random variable is a constant

Almost Sure Convergence

- Also known as **strong convergence**
- Formal definition:

Definition 3 (Almost sure convergence)

A random sequence $\{X_i\}$ converges almost surely to the random variable X if:

$$Prob(\lim_{n\to\infty} X_n = X) = 1$$

- *Interpretation:* events for which X_n does not converge to X have probability equal to 0
- Notation:

$$X_n \overset{a.s.}{\to} X$$

- Almost sure convergence implies convergence in probability

Sure Convergence

Definition 4 (Sure convergence)

A sequence $\{X_i\}$ of random variables defined on the same probability space (i.e. a random process) converges surely to X if:

$$\lim_{n \to \infty} X_n(\omega) = X(\omega),\ \forall \omega \in \Omega$$

where Ω is the sample space of the underlying probability space.

- Implies almost sure convergence
- However, there is really not much "value added" from using sure convergence over almost sure convergence, so it's very rarely used

Convergence in Mean

Definition 5 (Convergence in mean)

A random sequence $\{X_i\}$ converges in the rth mean to the random variable X if the rth absolute moments $\mathsf{E}(|X_n|^r)$ and $\mathsf{E}(|X|^r)$ exist and:

$$\lim_{n\to\infty} \mathsf{E}(|X_n - X|^r) = 0$$

- When $r = 1$, we say that we have **convergence in mean**
- When $r = 2$, we say that we have **convergence in mean square**
- Convergence in the rth mean implies convergence in probability

Review of OLS

The OLS Estimator

- Start from the matrix form of the regression relationship:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

  where $\mathbf{y}$ is an $(n \times 1)$ vector of values of the dependent variable; $\mathbf{X}$ is a $(n \times k)$ matrix of values of independent variables $x_1, \ldots, x_k$; $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of parameters; and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of stochastic disturbances
- Note that $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are population characteristics
- For each observation, we have:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

  where $\mathbf{x}_i'$ is the ith row of the $\mathbf{X}$ matrix

The OLS Estimator (2)

- From the latter we can express the disturbances as follows:

$$\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n$$

- Correspondingly, after we find the sample counterparts of population parameters, i.e. **b**, we can write an expression for the estimated regression residuals:

$$e_i = y_i - \mathbf{x}_i'\mathbf{b}, \quad i = 1, 2, \ldots, n$$

- The issue is how to choose $\mathbf{b}$ so that the regression line is as close as possible to all points
- We minimize the sum of squared residuals:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} e_i^2 = \min_{\mathbf{b}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b})^2$$

The OLS Estimator (3)

- This is the same as minimizing $\mathbf{e}'\mathbf{e}$:

$$\min_{\mathbf{b}} S(\mathbf{b}) = \min_{\mathbf{b}} (\mathbf{e}'\mathbf{e}) = \min_{\mathbf{b}} [(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})]$$

- Expand the RHS of this to get: Derivation

$$\min_{\mathbf{b}} S(\mathbf{b}) = \min_{\mathbf{b}} (\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b})$$

- Differentiate with respect to $\mathbf{b}$ and set the derivatives to zero (first-order condition, FOC):

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

- Since we start from a quadratic function which is convex, the FOC is also sufficient

The OLS Estimator (4)

- To formalize the latter, look at the second-order condition (SOC):

$$\frac{\partial^2 \mathbf{S}(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X}$$

- For a minimum, $\mathbf{X}'\mathbf{X}$ should be positive definite
- Take an arbitrary non-zero vector $\mathbf{c}$ and let $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$; then:

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^{n} v_i^2,$$

  where $\mathbf{v} = \mathbf{X}\mathbf{c}$
- $\mathbf{v}$ is non-zero, otherwise it would be a linear combination of the columns in $\mathbf{X}$, the latter implying that $\mathbf{X}$ does not have full column rank
- Therefore, $q > 0$, implying $\mathbf{X}'\mathbf{X}$ (and also $2\mathbf{X}'\mathbf{X}$) is positive definite

The OLS Estimator (5)

- Rearranging leads to

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

- If $(\mathbf{X}'\mathbf{X})^{-1}$ exists, we can multiply both sides by it
- The minimum is achieved therefore at:

$$\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}} \quad (*)$$

- Now take the residuals from the OLS regression:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

The OLS Estimator (6)

- Consider Application 3.2.2 from Greene's book
- We will replicate the final form of the investment equation
- The data and the R code are already provided for you on the course webpage

Algebraic Aspects of the OLS Solution

- Start from $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$
- Rearrange to get the **normal equations**:

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = -\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = -\mathbf{X}'\mathbf{e} = \mathbf{0}$$

- The latter means that for each column $\mathbf{x}_k$ of $\mathbf{X}$,

$$\mathbf{x}_k'\mathbf{e} = 0$$

- Assume that the first column of $\mathbf{X}$ is a column of 1s; denote it by $\mathbf{i}$
- This leads to three implications (see next two slides)

Algebraic Aspects of the OLS Solution (2)

1. *The LS residuals sum to zero.* Follows from:

$$\mathbf{x}_1'\mathbf{e} = \mathbf{i}'\mathbf{e} = \sum_{i=1}^{n} e_i = 0$$

2. *The regression hyperplane passes through the point of means of the data.* From the first normal equation we have

$$\sum_{i=1}^{n} e_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b}) = 0$$

Divide the latter by n to get:

$$\bar{y} = \bar{\mathbf{x}}'\mathbf{b}$$

Algebraic Aspects of the OLS Solution (3)

3. *The mean of the fitted values equals the mean of the actuals*. The fitted values are $\mathbf{x}_i'\mathbf{b}$, while the actual ones are y_i; the equality was already established in 2. ($\overline{y} = \overline{\mathbf{x}}'\mathbf{b}$)

Note: All the above are valid for regressions with an intercept term. If there is no intercept term, the three implications might not be valid.

Interesting Matrices Related to the OLS Estimator

- Substitute $(*)$ for $\mathbf{b}$:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{M}\mathbf{y}$$

where $\mathbf{M}$ is an $n \times n$ symmetric and idempotent matrix:

$$(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}')(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}') = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$$

- $\mathbf{M}$ is interpreted as the matrix that produces the vector of OLS residuals (**“the residual maker”**)
- It is easy to see that

$$\mathbf{M}\mathbf{X} = \mathbf{0},$$

i.e. regressing $\mathbf{X}$ on $\mathbf{X}$ produces a perfect fit, i.e. no residuals

Interesting Matrices Related to the OLS Estimator (2)

- The sample analogue of the regression relationship $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ can also be written as:
$$\mathbf{y} = \widehat{\mathbf{y}} + \mathbf{e}$$
- If we take the product $\widehat{\mathbf{y}}'\mathbf{e} = \mathbf{b}'\mathbf{X}'\mathbf{M}\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{M}'\mathbf{y} = (\mathbf{M}\mathbf{X}\mathbf{b})'\mathbf{y}$, then using the fact that $\mathbf{M}\mathbf{X} = \mathbf{0}$, we have that:
$$\widehat{\mathbf{y}}'\mathbf{e} = \mathbf{0},$$
i.e. $\hat{\mathbf{y}}$ and $\mathbf{e}$ are orthogonal to each other

Interesting Matrices Related to the OLS Estimator (3)

- We also have:

$$\widehat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = \mathbf{y} - \mathbf{M}\mathbf{y} = (\mathbf{I} - \mathbf{M})\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}$$

  $\mathbf{P}$ is a **projection matrix**: it projects $\mathbf{y}$ into the vector of fitted values $\widehat{\mathbf{y}}$
- We can show that $\mathbf{P}$ is also symmetric and idempotent
- Additionally:

$$\mathbf{PM} = \mathbf{MP} = \mathbf{0},$$

  i.e. $\mathbf{P}$ and $\mathbf{M}$ are orthogonal
- The projection of $\mathbf{X}$ is naturally $\mathbf{X}$, i.e.:

$$\mathbf{PX} = \mathbf{X}$$

- Combining results, we can write also:

$$\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y} = \textbf{projection + residual}$$

Spherical Disturbances

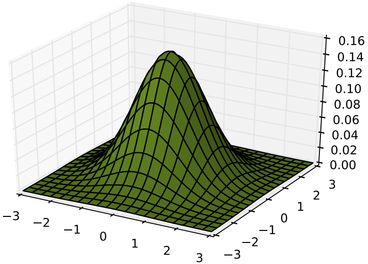
- Recall: two of the CLRM requirements are that the regression random disturbances are homoskedastic and are not autocorrelated
- Using the expectation operator and a bit of matrix notation, we can write the above in shorthand as follows:

$$\mathsf{E}(\varepsilon\varepsilon') = \sigma^2\mathbf{I}$$

- In other words, the variance-covariance matrix of the vector of disturbances ε should be diagonal with each diagonal element equalling σ^2

Spherical Disturbances (2)

- So, where does “spherical” come from?
- Take the graph of the bivariate standard normal distribution as an example:[1]



- If you intersect this surface with a plane parallel to the xy-plane, the intersection will be a circumference

[1] Recall that a probability distribution is a mathematical model of the variation of outcomes of a random variable.

Spherical Disturbances (3)

- In analytical terms, the equation:

$$f(\mathbf{x}) = c,$$

where $f(\mathbf{x}) = \dfrac{1}{2\pi}\exp\left(-\dfrac{\mathbf{x}'\mathbf{x}}{2}\right)$ is the bivariate standard normal pdf, is an equation of a circumference

- In the three- and higher-dimensional cases, the circumference becomes a sphere; therefore "spherical"

Finite-Sample Properties of $\mathbf{b}$: Variance

- Under the assumption of spherical disturbances, $\mathsf{Var}(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{I}$
- The conditional variance of $\mathbf{b}$ is:

$$\begin{aligned}\mathsf{Var}(\mathbf{b}|\mathbf{X}) &= \mathsf{E}[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'|\mathbf{X}] = \mathsf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \\ &= \boxed{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}}\end{aligned}$$

Finite-Sample Properties of $\mathbf{b}$: Variance (2)

- Example: In the simple (bivariate) regression case, the $\mathbf{X}$ matrix is of dimensions $(n \times 2)$ where the first column consists of 1s
- $\mathbf{X}'\mathbf{X}$ equals:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

- Its inverse multiplied by σ^2 is:

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\sigma^2}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

- The variance of the slope coefficient b is equal to the $(2,2)$ element of the matrix:

$$\text{Var}(b|\mathbf{X}) = \frac{\sigma^2 n}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sigma^2}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2}$$

- From this point onward it is easy to show that:

$$\text{Var}(b|\mathbf{X}) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Finite-Sample Properties of $\mathbf{b}$: Variance (3)

- In order to test hypotheses about $\boldsymbol{\beta}$, an estimate of the $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ matrix is required
- $\mathbf{X}$ is simply data, so only an estimate of σ^2 is needed
- We know that $\sigma^2 = \mathrm{E}(\varepsilon_i^2)$; also, e_i are estimates of ε_i
- It seems logical to use then

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

- However, e_i are imperfect estimators of ε_i as

$$e_i = y_i - \mathbf{x}_i'\mathbf{b} = y_i - \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\mathbf{b} = \varepsilon_i - \mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta})$$

- The far RHS suggests that $\widehat{\sigma}^2$ is a distorted estimate since it depends on the estimate of $\boldsymbol{\beta}$

Finite-Sample Properties of $\mathbf{b}$: Variance (4)

- Using the fact that $\mathbf{MX} = \mathbf{0}$, the OLS residuals are

$$\mathbf{e} = \mathbf{My} = \mathbf{M}(\mathbf{X}\beta + \varepsilon) = \mathbf{M}\varepsilon$$

- The estimator of σ^2 is based on the sum of squared residuals

$$\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}\varepsilon$$

- The expression $\mathbf{e}'\mathbf{e}$ is a quadratic form, and its expected value is

$$\mathsf{E}(\mathbf{e}'\mathbf{e}|\mathbf{X}) = \mathsf{E}(\varepsilon'\mathbf{M}\varepsilon|\mathbf{X})$$

- $\varepsilon'\mathbf{M}\varepsilon$ is a scalar, therefore $\varepsilon'\mathbf{M}\varepsilon = \mathrm{tr}(\varepsilon'\mathbf{M}\varepsilon)$

Finite-Sample Properties of $\mathbf{b}$: Variance (5)

- The trace of a product of square matrices remains the same irrespective of the permutation of product elements
- This can be transferred to a square matrix pre-multiplied and post-multiplied by a vector such as $\varepsilon'\mathbf{M}\varepsilon$,

$$\mathsf{E}[\mathrm{tr}(\varepsilon'\mathbf{M}\varepsilon)|\mathbf{X}] = \mathsf{E}[\mathrm{tr}(\mathbf{M}\varepsilon\varepsilon')|\mathbf{X}]$$

- The $\mathbf{M}$ matrix is a function of $\mathbf{X}$, therefore

$$\mathsf{E}[\mathrm{tr}(\mathbf{M}\varepsilon\varepsilon')|\mathbf{X}] = \mathrm{tr}(\mathbf{M}\mathsf{E}[\varepsilon\varepsilon'|\mathbf{X}]) = \mathrm{tr}(\mathbf{M}\sigma^2\mathbf{I}) = \sigma^2\mathrm{tr}(\mathbf{M})$$

- The trace of $\mathbf{M}$ is only needed here; recall that $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Finite-Sample Properties of $\mathbf{b}$: Variance (6)

- Then:

$$\begin{aligned}\mathrm{tr}(\mathbf{M}) &= \mathrm{tr}[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathrm{tr}[\mathbf{I}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \\ &= \mathrm{tr}(\mathbf{I}) - \mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n-K\end{aligned}$$

- Here we used the fact that $\mathbf{I}$ is $(n \times n)$ while $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ is $(K \times K)$ where K is the number of regressors (incl. the constant)
- It follows that
$$\mathsf{E}(\mathbf{e}'\mathbf{e}|\mathbf{X}) = \mathsf{E}(\varepsilon'\mathbf{M}\varepsilon|\mathbf{X}) = (n-K)\sigma^2$$
- If we use the sample estimator for σ^2, i.e. s^2, we get $\dfrac{n-K}{n}\sum_{i=1}^{n} e_i^2$ which is biased downwards

Finite-Sample Properties of $\mathbf{b}$: Variance (7)

- An unbiased estimator is therefore

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$$

- The standard error of the regression equals s
- The variance estimated from sample data is then

$$\text{Est.Var}[\mathbf{b}|\mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

- The diagonal elements of this matrix are the variances of the estimated individual regression coefficients, b_k
- The square roots of those elements correspondingly provide the **standard errors** of the estimators

The Gauss-Markov Theorem

Theorem 1 (Gauss-Markov Theorem)

The least squares estimator $\mathbf{b}$ *in the linear regression model is the minimum variance linear unbiased estimator of* $\boldsymbol{\beta}$*. For any vector of constants* $\mathbf{w}$*, the minimum-variance linear unbiased estimator of* $\mathbf{w}'\boldsymbol{\beta}$ *is* $\mathbf{w}'\mathbf{b}$

Finite-Sample Properties of $\mathbf{b}$: Unbiasedness

- According to the Gauss-Markov theorem, the OLS estimator $\mathbf{b}$ is BLU
- Start with:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

- From this, linearity with respect to $\boldsymbol{\varepsilon}$ is obvious
- To prove unbiasedness, take expectations conditional on $\mathbf{X}$

$$\mathsf{E}(\mathbf{b}|\mathbf{X}) = \boldsymbol{\beta} + \mathsf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}]$$

- The far RHS term equals $\mathbf{0}$, so $\mathsf{E}(\mathbf{b}|\mathbf{X}) = \boldsymbol{\beta}$
- Using the law of iterated expectations:

$$\mathsf{E}_{\mathbf{X}}[\mathsf{E}(\mathbf{b}|\mathbf{X})] = \mathsf{E}(\mathbf{b}) = \boldsymbol{\beta}$$

Finite-Sample Properties of $\mathbf{b}$: Efficiency

- To prove "best" (i.e. efficient/minimum-variance), take another arbitrary unbiased linear estimator of β:

$$\mathbf{b}_0 = \mathbf{C}\mathbf{y} = \mathbf{C}(\mathbf{X}\beta + \varepsilon) = \mathbf{C}\mathbf{X}\beta + \mathbf{C}\varepsilon$$

  where $\mathbf{C}$ is of dimensions $(k \times n)$
- Due to unbiasedness, we have:

$$\mathsf{E}(\mathbf{C}\mathbf{y}|\mathbf{X}) = \mathsf{E}[\mathbf{C}(\mathbf{X}\beta + \varepsilon)|\mathbf{X}] = \mathsf{E}[\mathbf{C}\mathbf{X}\beta + \mathbf{C}\varepsilon)|\mathbf{X}] = \beta$$

- From this follows that $\mathbf{C}\mathbf{X} = \mathbf{I}$
- Let $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so that $\mathbf{D}\mathbf{y} = \mathbf{b}_0 - \mathbf{b}$

Finite-Sample Properties of $\mathbf{b}$: Efficiency (2)

- The conditional variance of $\mathbf{b}_0$ is:

$$
\begin{aligned}
\text{Var}(\mathbf{b}_0|\mathbf{X}) &= \text{E}[(\mathbf{b}_0-\boldsymbol{\beta})(\mathbf{b}_0-\boldsymbol{\beta})'|\mathbf{X}] = \\
&= \text{E}[(\mathbf{CX}\boldsymbol{\beta}+\mathbf{C}\boldsymbol{\varepsilon}-\boldsymbol{\beta})(\mathbf{CX}\boldsymbol{\beta}+\mathbf{C}\boldsymbol{\varepsilon}-\boldsymbol{\beta})'|\mathbf{X}] = \\
&= \text{E}(\mathbf{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{C}'|\mathbf{X}) = \sigma^2\mathbf{CC}' = \\
&= \sigma^2\{[\mathbf{D}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{D}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'\} = \\
&= \sigma^2\{[\mathbf{D}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{D}'+\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\} = \\
&= \sigma^2[\mathbf{DD}'+\mathbf{DX}(\mathbf{X}'\mathbf{X})^{-1}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}+ \\
&+ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}']
\end{aligned}
$$

Finite-Sample Properties of $\mathbf{b}$: Efficiency (3)

- Using the fact that $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{CX} = \mathbf{I}$, we have:

$$\mathbf{DX} = \mathbf{CX} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I} - \mathbf{I} = \mathbf{0}$$

- Noting also that $\mathbf{X}'\mathbf{D}' = (\mathbf{DX})'$, the variance simplifies to:

$$\mathrm{Var}(\mathbf{b}_0|\mathbf{X}) = \sigma^2[\mathbf{DD}' + (\mathbf{X}'\mathbf{X})^{-1}] = \mathrm{Var}(\mathbf{b}|\mathbf{X}) + \sigma^2\mathbf{DD}'$$

- We can quickly show that $\mathbf{DD}'$ is positive semi-definite:

$$\mathbf{x}'\mathbf{DD}'\mathbf{x} = (\mathbf{D}'\mathbf{x})'\mathbf{D}'\mathbf{x} = \mathbf{z}'\mathbf{z} \geq 0$$

  (In the latter we also make use of the fact that $\mathbf{D}$ has more rows (n) than columns (k), therefore it cannot have full rank; in other words, there exists some $\mathbf{x} \neq \mathbf{0}$ for which $\mathbf{D}'\mathbf{x} = 0$.)

- The conclusion is that the variance of $\mathbf{b}_0$ is at least as large as that of $\mathbf{b}$

Finite-Sample Properties of $\mathbf{b}$: Stochastic Regressors

- What if $\mathbf{X}$ is stochastic (i.e. we have stochastic regressors)?
- We already have the conditional mean and variance
- We can use that to obtain the unconditional variance by averaging over all possible values of $\mathbf{X}$
- Using the variance decomposition formula, we have:

$$\mathrm{Var}(\mathbf{b}) = \mathrm{E}_{\mathbf{X}}[\mathrm{Var}(\mathbf{b}|\mathbf{X})] + \mathrm{Var}_{\mathbf{X}}[\mathrm{E}(\mathbf{b}|\mathbf{X})]$$

- The far right term is 0 since we have a constant vector in the square brackets (β)
- Therefore:

$$\mathrm{Var}(\mathbf{b}) = \mathrm{E}_{\mathbf{X}}[\mathrm{Var}(\mathbf{b}|\mathbf{X})] = \mathrm{E}_{\mathbf{X}}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2\mathrm{E}_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}]$$

- Because efficiency is valid for any value of $\mathbf{X}$, then it is valid also for an average of the possible values of $\mathbf{X}$

Finite-Sample Properties of $\mathbf{b}$: Normality

- All the discussion so far did not use any assumption on the distribution of ε
- In other words, the specification that was used is **semiparametric**
- Now recall $\mathbf{b}$ is a linear function of ε
- If ε is multivariate normal, then the conditional distribution of $\mathbf{b}$ is

$$\mathbf{b}|\mathbf{X} \sim \mathcal{N}[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$$

- For each element of $\mathbf{b}$ we have

$$b_k|\mathbf{X} \sim \mathcal{N}[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}]$$

Asymptotic Properties of $\mathbf{b}$: General Remarks

- Unbiasedness in finite samples is a useful indicator whether the estimator does not underestimate of overestimate *systematically* the true value
- However, it has significant limitations when it comes to estimation strategy:
  - Except in the linear regression model, the estimator is rarely unbiased
  - Unbiasedness in no way tells us that more information (a larger sample) is better than less
- The desirable property of an estimator is therefore the following: when sample size increases, it gets more accurate
- Thus, we look for **consistency** and not so much for unbiasedness

Asymptotic Properties of $\mathbf{b}$: Consistency

- Assume that the data generating mechanism for $\mathbf{X}$ is unspecified
- In particular, this means that $\mathbf{X}$ can contain both stochastic and deterministic vectors independent from ε
- Two critical assumptions are made:
    - $(\mathbf{x}_i, \varepsilon_i)$ is a sequence of i.i.d. *observations* (not variables)
    - Assume also that asymptotically the data behave as follows:

$$\operatorname*{plim}_{n\to\infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}$$

      where $\mathbf{Q}$ is a positive definite matrix
- It is clear that if $\mathbf{X}$ has full column rank, then $\frac{\mathbf{X}'\mathbf{X}}{n}$ is positive definite in a *specific* sample of $n \geq K$ observations
- The second critical assumption extends the above conclusion for a specific sample to *any* sample which has $n \geq K$ observations

Asymptotic Properties of $\mathbf{b}$: Consistency (2)

Theorem 2 (Rules for probability limits)

Let X_n *and* Y_n *be sequences of random variables such that* $\operatorname{plim} X_n = c$ *and* $\operatorname{plim} Y_n = d$*. Then:*

- $\operatorname{plim}(X_n + Y_n) = c + d$
- $\operatorname{plim}(X_n Y_n) = cd$
- $\operatorname{plim}(X_n / Y_n) = c/d$, *if* $d \neq 0$

If $\mathbf{W}_n$ *is a sequence of matrices of random variables and* $\operatorname{plim} \mathbf{W}_n = \mathbf{\Omega}$*, then*

$$\operatorname{plim} \mathbf{W}_n^{-1} = \mathbf{\Omega}^{-1}.$$

If $\mathbf{X}_n$ *and* $\mathbf{Y}_n$ *are sequences of matrices of random variables with* $\operatorname{plim} \mathbf{X}_n = \mathbf{A}$ *and* $\operatorname{plim} \mathbf{Y}_n = \mathbf{B}$ *then*

$$\operatorname{plim} \mathbf{X}_n \mathbf{Y}_n = \mathbf{AB}$$

Asymptotic Properties of $\mathbf{b}$: Consistency (3)

- From this theorem we can in particular infer that

$$\underset{n\to\infty}{\operatorname{plim}} \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} = \mathbf{Q}^{-1}$$

- The critical assumptions that were made might turn out to be too restrictive, especially in the case of trending or polynomial time series

Asymptotic Properties of $\mathbf{b}$: Consistency (4)

- The **Grenander assumptions** are a weaker alternative that includes most of the critical assumptions and ensures that the data matrix is “well behaved” in large samples:
  1. $\lim_{n\to\infty} d_{nk}^2 = \lim_{n\to\infty} \mathbf{x}_k'\mathbf{x}_k = +\infty$, where $\mathbf{x}_k$ is an arbitrary column of $\mathbf{X}$; interpretation: no variable will degenerate into a sequence of zeros
  2. $\lim_{n\to\infty} \frac{x_{ik}^2}{d_{nk}^2} = 0,\ \forall i$; interpretation: there will be no single observation that dominates the sum
  3. If $\mathbf{R}_n$ is the correlation matrix of the columns of $\mathbf{X}$ excluding the constant, then $\lim_{n\to\infty} \mathbf{R}_n = \mathbf{C}$, where $\mathbf{C}$ is positive definite; interpretation: $\mathbf{X}$ has full rank, i.e. there is no multicollinearity among regressors

Asymptotic Properties of $\mathbf{b}$: Consistency (5)

- Write the OLS estimator as:

$$\mathbf{b} = \beta + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$$

- If the inverse of $\mathbf{Q}$ exists, then:

$$\underset{n\to\infty}{\text{plim}}\,\mathbf{b} = \beta + \mathbf{Q}^{-1}\underset{n\to\infty}{\text{plim}}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$$

- Take the last term whose limit in probability we have to find:

$$\frac{\mathbf{X}'\varepsilon}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \overline{\mathbf{w}}$$

(In the above, $\mathbf{x}_i$ is the ith column of $\mathbf{X}$)

- So, we can rewrite:

$$\underset{n\to\infty}{\text{plim}}\,\mathbf{b} = \beta + \mathbf{Q}^{-1}\underset{n\to\infty}{\text{plim}}\,\overline{\mathbf{w}}$$

Asymptotic Properties of $\mathbf{b}$: Consistency (6)

- By assumption, in the CLRM the regressors are exogenous so (using the law of iterated expectations again):

$$\mathsf{E}(\mathbf{w}_i) = \mathsf{E}_{\mathbf{X}}[\mathsf{E}(\mathbf{w}_i|\mathbf{x}_i)] = \mathsf{E}_{\mathbf{X}}[\mathbf{x}_i \mathsf{E}(\varepsilon_i|\mathbf{x}_i)] = \mathbf{0}$$

- This implies:

$$\mathsf{E}(\overline{\mathbf{w}}) = \mathbf{0}$$

- With respect to the variance, use the variance decomposition formula:

$$\mathsf{Var}(\overline{\mathbf{w}}) = \mathsf{E}_{\mathbf{X}}[\mathsf{Var}(\overline{\mathbf{w}}|\mathbf{X})] + \mathsf{Var}_X[\mathsf{E}(\overline{\mathbf{w}}|\mathbf{X})] = \mathsf{E}_{\mathbf{X}}[\mathsf{Var}(\overline{\mathbf{w}}|\mathbf{X})] + \mathbf{0}$$

- We have then:

$$\begin{aligned} \mathsf{Var}(\overline{\mathbf{w}}|\mathbf{X}) &= \mathsf{E}(\overline{\mathbf{w}}\overline{\mathbf{w}}'|\mathbf{X}) = \mathsf{E}\left(\frac{\mathbf{X}'\varepsilon}{n}\frac{\varepsilon'\mathbf{X}}{n}|\mathbf{X}\right) = \frac{\mathbf{X}'}{n}\mathsf{E}(\varepsilon\varepsilon'|\mathbf{X})\frac{\mathbf{X}}{n} = \\ &= \left(\frac{\sigma^2}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) \end{aligned}$$

Asymptotic Properties of $\mathbf{b}$: Consistency (7)

- Taking the limit in probability of the latter results in:

$$\operatorname*{plim}_{n\to\infty} \mathrm{Var}(\overline{\mathbf{w}}) = 0 \cdot \mathbf{Q} = \mathbf{0}$$

- Using this result, we end up with:

$$\operatorname*{plim}_{n\to\infty} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta},$$

which establishes consistency of $\mathbf{b}$ under OLS

Asymptotic Properties of $\mathbf{b}$: Distribution

- Finally, we will consider the asymptotic distribution of $\mathbf{b}$
- Assume that in addition to the requirement that regressors and disturbances are uncorrelated, we have also independent observations
- Note that we do not make any assumption for normality of disturbances!
- Recall that we could also write the OLS estimator as:

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$$

- We can rearrange this in the following way:

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}}\right)$$

Asymptotic Properties of $\mathbf{b}$: Distribution (2)

- We know that $\operatorname*{plim}_{n\to\infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}$; therefore, we only have to find the limiting distribution of what remains, i.e.:

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon = \left(\frac{1}{\sqrt{n}}\right) \cdot n \cdot \underbrace{\left(\frac{\mathbf{X}'\varepsilon}{n}\right)}_{=\overline{\mathbf{w}}} = \sqrt{n}(\overline{\mathbf{w}} - \mathsf{E}(\overline{\mathbf{w}}))$$

- We will use the Lindenberg-Feller version of the Central Limit Theorem to this end

Asymptotic Properties of $\mathbf{b}$: Distribution (3)

Theorem 3 (Lindenberg-Feller Theorem)

Let $\{x_i\}$, $i = 1, 2, \ldots, n$ *be a sequence of independent random variables with means* $\mu_i < \infty$ *and variances* $\sigma_i^2 < \infty$. *Let also:*

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\mu_i \quad \textit{and} \quad \overline{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\sigma_i$$

If $\lim_{n\to\infty} \dfrac{\max(\sigma_i)}{n\overline{\sigma}_n} = 0$ *(i.e. the largest term of the sum does not dominate the average), and if* $\lim_{n\to\infty} \overline{\sigma}_n = \overline{\sigma} < \infty$, *then:*

$$\sqrt{n}(\overline{x}_n - \overline{\mu}_n) \xrightarrow{d} \mathcal{N}(0, \overline{\sigma})$$

Asymptotic Properties of $\mathbf{b}$: Distribution (4)

- Start with the fact that $\overline{\mathbf{w}}$ is the average of n independent random vectors[2] $\mathbf{w}_i = \mathbf{x}_i \varepsilon_i$, with:

$$\mathsf{E}(\mathbf{w}_i) = \mathbf{0}$$
$$\mathsf{Var}(\mathbf{w}_i) = \mathsf{Var}(\mathbf{x}_i \varepsilon_i) = \sigma^2 \mathsf{E}(\mathbf{x}_i \mathbf{x}_i') = \sigma^2 \mathbf{Q}_i$$

- The variance of $\sqrt{n}$ times the average of all $\mathbf{w}_i$ (using independence) is:

$$\mathsf{Var}(\sqrt{n}\,\overline{\mathbf{w}}) = \mathsf{Var}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_i\right) = \frac{1}{n}\sigma^2 \sum_{i=1}^{n} \mathbf{Q}_i = \sigma^2 \overline{\mathbf{Q}}_n$$

- If there is no $\mathbf{Q}_i$ which dominates the sum, then:

$$\lim_{n\to\infty} \sigma^2 \overline{\mathbf{Q}}_n = \sigma^2 \mathbf{Q}$$

[2] The Theorem is also valid in a multivariate context.

Asymptotic Properties of $\mathbf{b}$: Distribution (5)

- Now we apply the Lindenberg-Feller Theorem:

$$\sqrt{n}(\overline{\mathbf{w}} - \mathsf{E}(\overline{\mathbf{w}})) = \left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q})$$

- As a consequence:

$$\mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} \mathcal{N}(\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}\sigma^2\mathbf{Q}\mathbf{Q}^{-1})$$

or:

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$$

Asymptotic Properties of $\mathbf{b}$: Distribution (6)

Theorem 4

If $\{\varepsilon_i\}$ are independently distributed with mean 0 and variance σ^2 and the observations x_{ik} comply with the Grenander conditions, then:

$$\mathbf{b} \overset{asy}{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right)$$

- Recall that we made no normality assumption
- Yet we arrived at an asymptotically normal distribution for the OLS estimator
- This follows from the CLT variant that we applied
- $\mathbf{Q}^{-1}/n$ is estimated in practice with $(\mathbf{X}'\mathbf{X})^{-1}$
- σ^2 is estimated with $s^2 = \dfrac{\mathbf{e}'\mathbf{e}}{n-k}$

Asymptotic Properties of s^2: Consistency

- We will show that s^2 is a consistent estimator
- Start from the fact that the OLS residuals are produced as follows:

$$\mathbf{e} = \mathbf{M}\mathbf{y}$$

- We can write this also as:

$$\mathbf{e} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = \mathbf{M}\boldsymbol{\varepsilon}$$

(because $\mathbf{M}\mathbf{X} = \mathbf{0}$)

- Then:

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\varepsilon}$$

Asymptotic Properties of s^2: Consistency (2)

- But $\mathbf{M}$ is symmetric and idempotent, so:

$$\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}\varepsilon$$

- Thus:

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} = \frac{\varepsilon'\mathbf{M}\varepsilon}{n-k}$$

- Substitute $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for its equal $\mathbf{M}$ and expand:

$$s^2 = \frac{n}{n-k}\left[\frac{\varepsilon'\varepsilon}{n} - \left(\frac{\varepsilon'\mathbf{X}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)\right]$$

Asymptotic Properties of s^2: Consistency (3)

- The following limits are easy to see (some of them we established in the preceding discussion):

$$\lim_{n\to\infty}\frac{n}{n-k}=1$$
$$\text{plim}_{n\to\infty}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}=\mathbf{Q}^{-1}$$
$$\text{plim}_{n\to\infty}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)=\text{plim}_{n\to\infty}\left(\frac{\varepsilon'\mathbf{X}}{n}\right)=\mathbf{0}$$

- Using the rules for multiplication of limits in probability, we finally end up with the following problem:

$$\text{plim}_{n\to\infty}s^2=\text{plim}_{n\to\infty}\frac{\varepsilon'\varepsilon}{n}=\text{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2=\overline{\varepsilon^2}$$

Asymptotic Properties of s^2: Consistency (4)

- By assumption, ε_i^2 are independent
- Assume also that ε_i are identically distributed
- We will use the following two:

Theorem 5 (Khinchin's Weak Law of Large Numbers)

If x_i, $i = 1, \ldots, n$ is a random i.i.d. sample from a distribution with finite mean $\mathrm{E}(x_i) = \mu$, then:

$$\operatorname{plim}_{n\to\infty} \overline{x} = \mu$$

Theorem 6 (Markov's Strong Law of Large Numbers)

If $\{z_i\}$ is a sequence of independent random variables with $\mathrm{E}(z_i) = \mu_i < \infty$ and if for some $\delta > 0$, $\sum_{i=1}^{\infty} \mathrm{E}(|z_i - \mu_i|^{1+\delta}]/i^{1+\delta} < \infty$, then

$$\overline{z}_n - \overline{\mu}_n \overset{a.s.}{\to} 0$$

Asymptotic Properties of s^2: Consistency (5)

- The mean of the random variables ε_i^2 is finite:

$$\mathsf{E}(\varepsilon_i^2) = \sigma^2 < +\infty$$

- Take $\delta = 1$; then $\mathsf{E}(|z_i - \mu_i|^{1+\delta}]$ becomes:

$$\mathsf{E}(\varepsilon_i^2 - \sigma^2)^2 = \mathsf{E}(\varepsilon_i^4 - 2\varepsilon_i^2\sigma^2 + \sigma^4) = \mathsf{E}(\varepsilon_i^4) - \sigma^4$$

- Assume now that $\mathsf{E}(\varepsilon_i^4)$ is finite (quite an "easy" assumption) so that the RHS term above is finite

Asymptotic Properties of s^2: Consistency (6)

- Then what follows is that:

$$\text{plim}_{n\to\infty} s^2 = \sigma^2$$

and

$$\text{plim}_{n\to\infty} s^2 \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2 \mathbf{Q}^{-1}$$

- Finally, we have the appropriate estimator of the asymptotic covariance matrix of $\mathbf{b}$:

$$\text{Est.Asy. Var}(\mathbf{b}) = s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Interval Estimation

Interval Estimation: General Considerations

- Objective: to present the best parameter estimate with an explicit expression of the estimate uncertainty, i.e.:

$$\widehat{\theta} \pm \text{sampling variability}$$

- If we want complete certainty, then we should take $\widehat{\theta} \pm \infty$: Not informative!
- If we stick to $\widehat{\theta} \pm 0$, this is also not desirable since the probability of being 100% precise is 0
- The point is to select some $\alpha \in (0, 1)$ so that a $(100 - \alpha)$% confidence interval is constructed

Forming a Confidence Interval for a Coefficient

- Assume that the disturbances ε_i are normally distributed
- Then for any element of $\mathbf{b}$,

$$b_k \sim \mathcal{N}(\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk})$$

- Standardize the latter to get

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim \mathcal{N}(0,1)$$

- Form a 95% confidence interval for z_k:

$$Prob[-1.96 \leq z_k \leq 1.96] = 0.95$$

Forming a Confidence Interval for a Coefficient (2)

- Return to the definition of z_k to rewrite the latter as follows:

$$Prob\left[b_k - 1.96\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}} \leq \beta_k \leq b_k + 1.96\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}\right] =$$

$$= 0.95$$

- However, the true variance is unknown; therefore, the ratio is modified so that the true variance is replaced by s^2 (the estimated regression variance):

$$t_k = \frac{b_k - \beta_k}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_{n-K}$$

- The latter is used to construct confidence intervals and test hypotheses about the elements of $\boldsymbol{\beta}$

Forming a Confidence Interval for a Coefficient (3)

- In particular, a confidence interval for β_k would be

$$Prob\left[b_k - t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}} \leq \beta_k \leq b_k + t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}}\right] =$$

$$= 1 - \alpha$$

where $S^{kk} = (\mathbf{X}'\mathbf{X})^{-1}_{kk}$ and $t_{(1-\alpha/2),[n-K]}$ is the corresponding critical value from the t distribution

Forming a Confidence Interval for a Coefficient (4)

- If the random disturbances ε are not normally distributed, then the result that the following statistic has a limiting standard normal distribution is used:

$$z_k = \frac{\sqrt{n}(b_k - \beta_k)}{\sqrt{\sigma^2 \mathbf{Q}^{kk}}}$$

where $\mathbf{Q} = \left[\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)\right]^{-1}$

- The Slutsky theorem:

Theorem 7 (Slutsky)

If $g(x_n)$ is a continuous function of x_n but not of n, then

$$\text{plim}\, g(x_n) = g(\text{plim}\, x_n)$$

Forming a Confidence Interval for a Coefficient (5)

- This theorem allows us to replace σ^2 in the formula with s^2 (a consistent estimator)
- The statistic modified this way still has the standard normal limiting distribution
- Since for a relatively large number of degrees of freedom (which also means relatively large sample size) the t distribution is indistinguishable from the normal one, the confidence interval would be

$$Prob\left[b_k - z_{(1-\alpha/2)}\sqrt{\text{Est.Asy.Var}(b_k)} \le \beta_k \le b_k + \right.$$

$$\left. + z_{(1-\alpha/2)}\sqrt{\text{Est.Asy.Var}(b_k)}\right] =$$

$$= 1 - \alpha$$

References

References

- Greene, W. (2013): *Econometric Analysis*, Pearson, 7th edn., ch. 3-4
- Davidson, R. and J. MacKinnon (2003): *Econometric Theory and Methods*, Oxford University Press, ch. 3
- Judge, G., W. Griffiths, R. Hill and T.-C. Lee (1980): *The Theory and Practice of Econometrics*, Wiley, 2nd edn., ch. 2

Extras

Derivation

Back

$$
\begin{aligned}
(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) &= (\mathbf{y}'-\mathbf{b}'\mathbf{X}')(\mathbf{y}-\mathbf{X}\mathbf{b}) = \\
&= \mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}\mathbf{b}-\mathbf{b}'\mathbf{X}'\mathbf{y}-\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}
\end{aligned}
$$

- Now we only need to show that $\mathbf{y}'\mathbf{X}\mathbf{b} = \mathbf{b}'\mathbf{X}'\mathbf{y}$
- Note that $\mathbf{y}$ is $(n \times 1)$, $\mathbf{X}$ is $(n \times k)$, and $\mathbf{b}$ is $(k \times 1)$
- Therefore any of the two will be (1×1), i.e. a scalar
- Since any of the two is the transpose of the other one, and given that those are scalars, the two expressions are equal